Symposium on Assessing Agency, Moral and Otherwise: Beyond the Machine Question

In conjunction with the 2018 Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2018)

4th April 2018

BOREALIS AI
RBC Institute for Research

Rilkean Memories for a Robot

Antonio Chella

Abstract. The paper discusses the role of Rilkean memories, recently introduced by Rowlands, in the building of the autobiographic self of a robot.

1 INTRODUCTION

It has been debated about the characteristics for an agent to be considered morally responsible for her actions. A generally recognized characteristic for moral agency is the capability for the agent to have a sense of self. According to this line of thinking, it has been debated whether a robot could ever be a morally responsible agent (see Gunkel [1], 46, for a discussion).
With the term "robot" we consider a humanoid robot, i.e., a mechanical entity with a human-like body shape, equipped with sensors like cameras, lasers, sonars, and with actuators like arms and legs, all controlled by a complex software system.
There is a long progression from a situation-action robot to a robot with a sense of self. In facts, the self of the robot is not something that is "uploaded," but, as in the humans, it develops in years after many interactions among the body of the robot, its control system, the users, the external environment, other robots and so on.
The current studies on the robot self essentially take into account the role of some model of episodic memory implemented in software by employing internal model methods or machine learning methods. However, the main role of the robot body in the building of the robot self is largely unexplored.
The paper claims that the development of the self of the robot, which is a complex issue involving several aspects, cannot ignore the aspects related with the body of the robot operating in the real world.

2 RILKEAN MEMORIES

Rilkean memories are a kind of autobiographical memories recently discussed by Rowlands [2,3], who took inspiration from the novel "The Notebooks of Malte Laurids Brigge" by Rainer Maria Rilke. Rilkean memories are:

> these memories that have become "blood," "glance and gesture," "nameless and no longer to be distinguished from ourselves" ([3], 54).

A Rilkean memory of an episode of the life of a person is related to the trace of that event left in the whole body of the individual, and not only in her brain.

[1] University of Palermo, Italy, email: antonio.chella@unipa.it

Rowlands discusses in detail how "this form of memory is typically, embodied and embedded; "it is a form of involuntary, autobiographical memory that is neither implicit nor explicit, neither declarative nor procedural, neither episodic nor semantic, and not Freudian." ([2], 141).
Rowlands points out that Rilkean memories of a person are responsible for the *style* of that person. An example is the motion style of the person: a person habit may be to walk on the left side of a path because of a traumatic episode during her life. The person may not be able to remember the traumatic episode explicitly, but the event entered into her blood, i.e., it becomes a part of her style.
Then, a Rilkean memory derives from episodic memory as a transformation of the act of remembering an episode in a behavioral and bodily disposition, even when the content of the episode is lost ([3], 73).
The whole self of a person is made up by different kind of memories, but, according to Rowlands, there is a strong relationship between the self of a person and the Rilkean memories. In facts, different people, with different histories and different episodes occurring in their life, acquire a distinct and recognizable personal style that survives even when the memories of the episodes occurring in the life of the person generating the memories, are lost. Then, the unity and persistence conditions that identify a person among the others is her style: "Rilkean memories … play a crucial role in holding the self together, in the face of certain well-documented facts." ([2], 154).

3 ROBOT RILKEAN MEMORIES

We claim that Rilkean memories may be considered even for robots, thus allowing it to build a robot self and to perform operations according to the acquired personal style.
In facts, the operation style of a robot is slowly acquired by the robot itself. The style does not depend uniquely on the software structures of its control system, as the outcomes of suitable neural networks storing the past interactions of the robot, but instead on the complex intermixing between the body of the robot and the software controlling the robot.
Consider, the NAO, a small humanoid robot built by SoftBank Robotics (previously Aldebaran) and commonly adopted in many research labs.
The RoboticsLab of the University of Palermo acquired two brand new NAOs in 2010. They were at the beginning mostly two instances of the same robot, and they were both capable of performing the same tasks in the very same ways. They had the same limitations due to their body constraints and due to their software characteristics.
The two brand new robots were employed for separate research projects with distinct purposes. As years went by, because of

their employment in different projects, the two NAO slightly changed their morphology in different ways. After performing and repeating several different actions, their motors started not to work correctly in idiosyncratic ways; their joints motions were not smooth enough and many bumps occurred in various parts of their bodies.

To clarify this point, let us consider the event of uploading a software program on one NAO that, for some reasons, caused the NAO arm to bump against the wall and thus generating a permanent malfunction of a robot joint. Then, a Rilkean memory of this episode occurs: even when the software program that allowed the robot to bump against the wall has been modified or cancelled, the bumping episode has been transformed in the malfunction of the joint. Thus, even if the robot is unable to access the episode because its log file has been cancelled, the episode has been transformed in a Rilkean memory of the robot.

It also happened that, because of several traumatic events, a part of one robot broke and it has been substituted and renewed. The substitution of a NAO body part is a Rilkean memory of the sequence of events that allowed the previous body part to be broken.

Then, as the two robots have been employed in different experiments and for different goals, after so many years they now have slightly different bodies.

Let us consider for example the case of one of the two robots that shows a problem with the motor controlling the left arm, i.e., a Rilkean memory of a previous bumping event. The control system of this NAO must take into account this motion constraint to generate a suitable plan of operations that minimize the use of the impaired arm. The other robot does not present any problem with the left arm, but instead with the right leg, because of a different traumatic event. Similarly, the control system of the other robot generates a motion plan that minimizes the movements of the damaged limb.

Other traumatic events occurred during the operational life of a robot are transformed in the excessive warm-up of motors, the critical duration of batteries, the limitations in the motions of joints, the malfunctions of one camera, and so on.

Therefore, in all of these cases, the control systems of the robots have to take care of these incurring different Rilkean memories. For example, one of the two NAOs has a tendency of an excessive warm-up of a motor because of a previous episode of an excessive employment of this motor. This NAO must interrupt its actions quite often, and thus this robot performs its task in a characteristic rough and fragmented way. Another example is related with the cameras of the robot: the NAO has two cameras mounted on top and on the bottom of the head. One of the two NAOs has one camera broken because of a crash event, and then this robot has to move the head in a peculiar way to compensate the malfunction, which is a Rilkean memory of the head crash.

It should be noticed that the problems occurring with a real robot operating in real environment are difficult to simulate by employing a robot simulator. Therefore, the internal simulator method, which is typically adopted in the design of the software system control of a robot, is not able to deal with the Rilkean memories.

After years of operations, then two different performance styles emerge, that tightly depend on the intricate intermixing between the hardware, the software and the biographies of the two NAO.

When the two robots are reunited together and involved in the same task, the controller of each robot must take into account the different software capabilities and hardware constraints to generate two different plans for the two robots, for the very same task. Then, the two robots perform the same task in two different ways, according to their style acquired during their operational lives.

Therefore, the two robots earned different memories of past episodes of their lives and different functional styles. The bodies of the two robots summarize the autobiographic selves of the robots and their acquired different personalities.

4 CONCLUSIONS

Robot Rilkean memories are the traces of the occurring episodes left in the whole body of the robot even when the memories of the episodes is lost. In the current literature, the discussions about the self of a robot typically consider the software control system of the robot, while the role of the body has been of limited interest. Therefore, we maintain that a robot, after years of operations, acquires its functioning style, which is unique and different from the style of the other robots, even if from the same factory.

It should be noticed that the employment of robot simulators, which is a typical strategy in robot software design, does not take into account all the real-world issues related with Rilkean memories, as the degradation of the operations of the joints, of the camera, the motor because of occurring episodes. Then, the body of the robot is one of the main aspects of the autobiographic self of the robot.

ACKNOWLEDGEMENTS

The author wants to thank Mark Rowlands and Riccardo Manzotti for their comments on previous versions of this paper.

REFERENCES

[1] D.J. Gunkel. (2012). *The Machine Question*. MIT Press, Cambridge, MA.

[2] M. Rowlands. (2015). Rilkean Memory, in: *The Southern Journal of Philosophy*, **53**, 141–154.

[3] M. Rowlands. (2016). *Memory and the Self: Phenomenology, Science and Autobiography*. Oxford University Press, Oxford, UK.

Are the notions of agency and responsibility relevant to questions about machine ethics?

Bryony Pierce[1]

Abstract. Discussions of machine ethics that focus on the capacity of artificial agents to act autonomously and on whether such agents should be held morally responsible for their actions should be abandoned in favour of an approach that prioritises concerns about whether intelligent machines might instead qualify as moral patients. When there is no agreement on whether human agents have free will or moral responsibility, or what this consists in, introducing such theory-laden notions is unhelpful, especially when the avoidance of suffering and the correct ascription of rights are recognised as being more important than who can or should be held accountable when harm is done. I set aside the notion of (moral) agency, focusing instead on the conditions necessary for moral patiency, arguing that this depends on the capacity for conscious affective experience and rejecting functional accounts of emotion that fail to incorporate the function of the qualitative character of affective experience.

1 INTRODUCTION

When considering attributing agency of any kind to machines with artificial intelligence (AI), or supposing that AI instantiated in autonomous robots or other systems could be held responsible for decisions and actions – in the future, if not now – theorists are making assumptions on the basis of widely accepted views of the characteristics possessed by so-called rational and moral agents. They are, explicitly or implicitly, considering the extent to which such machines can be similar to human agents, thus envisaged, in one broad area of functioning.

2 FREE WILL

One factor viewed as central, whether explicitly or implicitly, when speaking of moral agency and responsibility, is free will – a commonly expressed concern is that without free will, there can be no moral responsibility. But there is no consensus on whether free will exists, or, amongst those who believe it does exist, on the conditions under which it is possible to act freely or what it means to exercise free will.

A major area of disagreement is the question of whether free will is compatible with determinism and/or quantum theory. If determinism is true and human behaviour is part of a causal chain in which the causes of action are ultimately external to the self, yet 'causal links are not enough for control', as Dennett claims [3, p. 72], the relevant problem would be that of whether machines can act for reasons, processing *abstracta* with more than syntactic content – in order to avoid the symbol grounding problem [5]. If, on the other hand, freely performed actions are made possible by the occurrence of random or probabilistic processes (see [4]), only AI using quantum computers might be thought to have the relevant capacities necessary for freedom of action and thus, potentially, moral responsibility.

Approaching machine ethics from this perspective will inevitably result in widespread disagreement on the criteria for moral agency and moral responsibility, arising largely from the diversity of underlying views on freedom of the will, some of which may be tacitly and unquestioningly endorsed. Moral agency and moral responsibility dependent upon the theory-laden notion of freedom of the will are therefore complex and controversial concepts. Furthermore, as there is no consensus on whether even humans are morally responsible for their actions, trying to justify a distinction between free human agency and machine behaviour, or between the behaviour of different types of AI, on the basis of these putative capacities is not a useful exercise.

3 FROM MORAL AGENCY TO MORAL PATIENCY

I therefore question the appropriateness of attempting to draw a comparison between human and AI action in terms of moral agency, arguing that the concepts of moral agency and moral responsibility are not relevant in the field of machine ethics. Instead of questioning whether or why intelligent machines might at some point have moral agency and thus potentially be held responsible for their actions, I argue that it is important to discuss under what circumstances, if any, we may need to show concern for certain machines as something more than material objects, the harming of which might cause harm or distress only to other humans. When, if at all, could an artificial agent be deemed to have moral patiency and to have rights on this basis, regardless of the question of responsibility?

Once we start by viewing artificial agents as potentially having moral patiency rather than moral agency, it becomes significant that, although the degree to which human agents are deemed responsible for their actions affects society's response (they may be referred for psychiatric treatment rather than sent to prison, for example), the standard response to unacceptable behaviour by artificial agents would typically be discontinuation or reprogramming, and this decision would not depend on whether they had a capacity for moral responsibility or, if so, were judged to be morally responsible for their actions. As Arbib says: "when a machine 'goes wrong,' there should be maintenance routines to fix it that would be very different from either the medical treatment or penal servitude applied to humans" [1, p. 372].

In machine ethics, we need to go not only beyond the machine question: whether machines can be agents, but beyond what I will call *the agency question*, which is the question of what qualifies any entity, artificial or otherwise, as an agent or moral agent. I will argue that the agency question is the real red herring

[1] Dept. of Philosophy, Univ. of Bristol, BS, UK. Email: bryony.pierce@bristol.ac.uk.

in debates about AI, personhood and ethics. Rather than considering what is required for (moral) agency, I set aside the notion of agency – I have argued elsewhere that all action is a kind of reaction or higher-order reaction [6] – claiming that ethical questions in the field of AI, as elsewhere, need not be concerned with degrees of autonomy or the nature of responsibility and should focus instead on moral patiency and the need for measures to protect sentient beings, whether human or artefactual, from direct or indirect harm.

This focus on moral patiency results in an asymmetry where there can be two broad categories of intelligent human or artificial entity: (a) those capable of knowingly causing harm and of experiencing events or states of affairs as harmful or bad in some way, for themselves or others, and (b) those capable of causing harm merely as instruments of human individuals or collectives and incapable of experiencing anything as harmful or bad. The key difference between these two kinds of intelligent entity is that only the former has the capacity for conscious affective responses, without which, I have argued, harm and suffering can have no subjectively meaningful semantic content. This is because judgements about good and bad, which are necessary for mastery of moral concepts (as well as concepts relating to prudential concern for oneself), are relative to values grounded in "affective responses to actual and potential states of affairs", with "[a]wareness of the qualitative nature of these responses [depending] on the what-it's-likeness of conscious experience" [7, p. 81]. It is this capacity for experiencing suffering, or pleasure, that makes conscious beings objects of moral concern, in that it is wrong to cause them to suffer, and allows those with sufficient reasoning ability to recognise the need for concern for other conscious entities in such a way that it is pragmatically valuable to hold them (i.e., those with this ability) accountable, if not morally responsible, if they cause harm to others.[2]

4 CONDITIONS FOR MORAL PATIENCY

An artificial agent may have the ability to reason from premises from some external source and could then reach the conclusion that it should avoid causing harm in a range of situations, but without affective responses it has no access to information that is subjectively meaningful upon which to base its decisions, so cannot act morally or immorally in its own right, only as an agent of its programmers or other users, blindly following instructions. Giving an intelligent machine the ability to learn and draw conclusions of its own that are not predicted or predictable by its programmers does not alter the fact that the machine is entirely reliant on whatever values, goals and principles were originally programmed into it. Without conscious affective responses, the machine lacks moral patiency, as it cannot experience suffering. It might still meet some minimal criteria for moral accountability, in that it might be able to act ethically to varying extents, and for others to respond accordingly might be useful, but being subject to the moral judgement of others is not sufficient for moral patiency in the sense of being an object of moral concern, which I claim could only be the case if the machine were capable of conscious affective experience that made it possible for it to suffer.

It might be argued that there could be cases of partial moral patiency, in either non-conscious or conscious artificial agents. I will consider both types:

> (1) A non-conscious artificial agent, however intelligent, would not qualify as a moral patient, even in some minimal sense, in my view, because it would not be capable of suffering. Even a non-conscious artificial agent that could be said to have goals and to desire their attainment, in the sense of having the function of evaluating certain goals as worth pursuing, would not be harmed by the thwarting of its goals, which might be seen as a functional equivalent of suffering and thus as justifying an ascription of partial moral patiency, because its goals and reasons for action would be grounded externally: "in the qualitative character of the conscious affective experience of their programmers or users" so would lack subjective meaningfulness [8].
>
> (2) In the case of a conscious artificial agent, if we assume that artificial agents either can or might one day be conscious, partial moral patiency would require the capacity for at least some suffering, and if this requirement were fulfilled, e.g., by the agent's having conscious affective responses,[3] the artificial agent would qualify as a moral patient *tout court*: the capacity for suffering, however minimal, would be sufficient for moral patiency.

There are doubtless other ways in which it might be argued that artificial agents might acquire partial moral patiency, some based on premises inconsistent with this account, but I take it that all cases would fall into one of the above two categories.

5 EMOTION AND SIMULATED EMOTION

Whereas the machine's programmed values (machine code) could be likened to the set of heritable characteristics (genetic code) that provides conscious beings with an initial basis for their values, the inability to experience suffering or to experience information as subjectively meaningful marks a distinction between non-conscious machines and conscious entities that cannot be treated in a similarly reductive manner. I am not saying that we need to remember that robots are merely programmed complex machines that cannot help doing what they do, but that we can view humans similarly as merely evolved complex living machines, reacting to situations in set ways over which they ultimately also have no conscious control.[4] In the case of humans, though, there is a capacity for

[2] By accountable, I mean that they can be called upon to explain their actions and expect to suffer certain consequences if they have caused harm or suffering without good reason; if morally responsible, we could, in addition, legitimately blame them for their wrongdoing and they would be deserving of whatever sanctions were appropriate in the particular circumstances.

[3] Consciousness, as I understand it, incorporates the capacity for affective responses; I argue that the function of consciousness is to act as an interface between cognition and emotion [7].

[4] This view, in abandoning the notion of moral responsibility in favour of the weaker *accountability* along with a requirement that those who harm others face the consequences of their actions, for pragmatic social reasons, has implications for ethical discourse more generally that I am willing to embrace, but which are not directly relevant to my argument in this paper.

affective responses – feelings – and the status of our feelings has value to us and thus to others who experience affective responses, in turn, to our circumstances. This reciprocal relation creates a social world, into which machines are gradually being welcomed, increasingly as proxies for human beings. Artificial agents are learning to simulate emotions and behave in a way that aims to show empathy through verbal responses and facial expressions, but behavioural complexity of this kind, devoid of conscious experience, cannot endow machines with either subjectively meaningful emotions or moral patiency.

I reject information-processing-based functional accounts of emotion, such as that outlined by Arbib [1, p. 376], who likens motivation to "biases which favor one strategy group over another" and emotion to "the way in which these biases interact with more subtle computations", or Sloman, Chrisley and Scheutz's reduction of emotion to "actual or potential disturbance of normal processing" [9, p. 230]. Sloman et al go on to claim that, in contrast to shallow models, their "architecture-based notions would allow people (or robots) to have joy, fear, anguish, despair, and relief despite lacking any normal way of expressing them" [9, p. 233]. Although there may be functional parallels that are indeed captured in such accounts, they fail to incorporate what I see as the equally important functional role of grounding meaning, values and reasons for action in the qualitative character of conscious affective experience. Robots, such as Softbank, Aldebaran and Yoshimoto Robotics Laboratory's 'Pepper', which its developers claim has emotions, merely simulate empathy and a subset of emotions and respond to emotional behaviour in order to satisfy consumer needs for robots capable of [simulating] social interaction. If robots were to be developed with emotions with the same functional roles – including grounding in the qualitative character of affective experience – as those of humans, the full details of which I will not go into here, attributing moral responsibility might be difficult (robots would have been programmed to develop along certain lines and arguing that they had free will, however great their autonomy in terms of their ability to learn independently and adapt their behaviour subsequently, would be controversial), but their capacity to suffer would grant them moral patiency.

People respond to machines, including those that they think of as completely unresponsive, as though they were agents, because humans are disposed to identify agents and to err on the side of seeing agency where there is none. This tendency to over-attribute agency is thought to be an extension of the evolutionarily adaptive practice of attributing agency to other humans: 'In so far as a non-human entity exhibits […] human-like features or behavioural cues, the psychological explanatory framework may become overextended to it as well' [2, p. 238]. So, people shout at vending machines, blame their satnavs for taking them the wrong way, behave as if their toasters have deliberately burnt the toast, and so on, not because they believe these machines are conscious or morally responsible, but because they have emotional responses to them that are similar to those they have when humans thwart them, which they are in the habit of expressing. A robot that is humanoid or a computer that speaks or interacts in other ways will perhaps be even more likely to produce this kind of response, if it stands in some kind of pseudo-social relation to its user, but my view is that, as things stand, the impression of interacting with a conscious entity would be illusory.

6 CONCLUSIONS

I have argued that machines that are not conscious and have no affective responses lack moral patiency, although our treatment of them might affect other conscious entities who are objects of moral concern. We may cause indirect harm to others by harming machines upon which they depend, for example, but artefactual entities lacking consciousness have no moral or other rights of their own. Entities that lack moral patiency, because they lack conscious affective responses, and have no meaningful access to the semantics of moral discourse – even ones that might, one day, process complex information about moral matters and produce well-informed, unbiased solutions to ethical questions – cannot be held morally responsible for their actions, and we should not even be asking whether they can be. Artefactual entities with artificial intelligence that are structured in such a way that they can experience affective responses, rather than merely simulating emotions, should, however, be accorded the status of moral patiency, regardless of whether they are thought to be moral agents or morally responsible for their actions. Agency and responsibility are theory-laden concepts that can be avoided in the field of machine ethics by shifting the focus to the question of moral patiency and the role of consciousness. Questions about moral agency and responsibility then become irrelevant.

REFERENCES

[1] Arbib, M. 'Beware the passionate robot' In Fellous, J-M & Arbib, M. (Eds.) *Who Needs Emotions? The Brain Meets the Robot* New York, Oxford University Press: 333-383 (2005).

[2] Csibra, G., Gergely, G., Bíró, S., Koós, O. & Brockbank, M. 'Goal attribution without agency cues: the perception of 'pure reason' in infancy', *Cognition* 72: 237-267 (1999).

[3] Dennett, D. *Elbow Room*, Oxford: Clarendon Press (1984).

[4] Hameroff, S.R. & Penrose, R. 'Orchestrated Reduction Of Quantum Coherence In Brain Microtubules: A Model For Consciousness?', in Hameroff, S.R., Kaszniak, A.W. & Scott, A.C. [Eds.], *Toward a Science of Consciousness - The First Tucson Discussions and Debates*, Cambridge, Massachusetts: MIT Press: 507-540 (1996).

[5] Harnad, S. 'The symbol grounding problem', *Physica*, D42: 335-346 (1990).

[6] Pierce, B. 'Is the concept of rational agency coherent?' *Philosophical Writings*, 33: 5-18 (2006).

[7] — 'Is the function of consciousness to act as an interface?' In F. Paglieri (Ed.), *Consciousness in interaction: the role of the natural and social context in shaping consciousness*, Amsterdam: John Benjamins: 73-88 (2012).

[8] — 'How are robots' reasons for action grounded?' (under review).

[9] Sloman, A, Chrisley, R. & Scheutz, M. 'The Architectural Basis of Affective States and Processes' In Fellous, J-M & Arbib, M. (Eds.) *Who Needs Emotions? The Brain Meets the Robot* New York, Oxford University Press: 203-244 (2005).

Machine Agency, Moral Relevance, and Moral Agency

John Preston[1]

Abstract. Machines, including robots, have always been agents and are becoming increasingly autonomous agents. But autonomy isn't sufficient for moral agency. Machines can be described in intentional terms, using what Dennett calls 'the intentional stance'. This doesn't yet make them fully intentional agents, for as yet we only apply certain aspects or parts of the intentional stance to machines.

There's no reason to think that we are (yet) developing genuinely moral machine or robot agents. There's a specific feature of decisions about the most morally weighty issues which means that we won't rightly think of *them* as having done the moral thinking required.

But moral *patiency*, as we might call it, matters, too. The extent to which artefacts will be credited with moral agency will also be affected by the extent to which we think of them as capable of genuine suffering. Until we are willing to credit machines and other robot agents with the capacity for thought, intention *and* suffering, we won't really think of them as moral agents.

Two matters that have been discussed in the literature here, the moral relevance of machines, and the 'neutrality thesis', are red herrings. The real question is whether machines can be credited with *moral* responsibility. Moral agency is a precondition for moral responsibility. Machines *are* morally relevant, and they are so at least partly in virtue of their being agents. But this doesn't make them moral agents. And the fact that the neutrality thesis is unacceptable doesn't indicate otherwise.

As our technology develops, we may begin to think of machines as genuinely *doing* some of the psychological things we now think of them as incapable of doing. But we won't think of them in moral terms until we come to think of them as *thinking about moral issues*, and as *knowing what's right and what's wrong* as a result of such thinking. We can't tell whether that day will come.

1 AGENCY, AND AGENTS

What is an agent? The bar is exceptionally low. An agent is not merely anything that acts, that is, anything that *does* something (usually *to* something else), but rather anything that *can* act, anything that *can* do something. So almost everything is an agent. Agency is, if not absolutely ubiquitous, then very widespread indeed.

There's nothing problematic about the notion of *inanimate* agents. We already speak not only of human agents but also of chemical agents, chemical warfare agents, biological agents, weathering agents, rinsing agents, etc. Social theorists of the 'actor-network' school (Bruno Latour et al.) are therefore not wrong to include non-human agents such as scallops, and electronic door-closers, in their networks (see [1,2,3]). In fact, all sorts of things, including phenomena that hardly make the grade as 'things', can be agents.

The existence of agents does not depend on the existence of humans. Even if humans had never existed, acids and alkali, for example, would still act, and there would still be chemical agents, weathering agents, etc. When we speak of such inanimate substances (stuffs) being agents, we have in mind their *tendency* or *disposition* to affect other things in certain specific ways, a tendency or disposition that may be triggered by certain circumstances. But the circumstances in question need involve no animate being. And agency can be as *basic* as one thing sitting on top of another – the thing doing the sitting on top of is exercising a kind of agency.

Agency contrasts with what we might call 'patiency', that is, being a patient, being something that something is *done to*. Obviously, being an agent isn't incompatible with being a 'patient' in this sense – it's quite possible for something to do things *and* to have things done to them, even at the same time. (Anything, call it A, sitting on top of anything else, B, as well as being an agent, is a patient in virtue of its being supported by B). So when I speak of a *contrast* between being an agent and being a patient, I mean only that one can contrast these two aspects of things. The agent/patient distinction is of no use in characterising the extension of the term 'agent', since almost everything is an agent, and almost everything is also a patient.

2 MACHINES AS AGENTS

Robots and other machines easily make the grade as agents, since as long as they function at all they're always capable of *doing* something, even if they're not doing it at that very moment.

Roboticists won't take much comfort in this conclusion, of course. For them, the fact that some acid in a bucket standing next to their latest robot creation is just as good an example of an agent as their robot itself is, may well be thought to devalue its claim to being an agent. This shows that *mere* agency is not the issue. Rather, we must move on to find some kind (or kinds) of agency that robots might be thought to have but which mere things like stones and mere stuffs like acids definitely don't have. Only such a kind of agency can be important, that is, worth having and worth striving to inculcate in robots. But what kind (or kinds) of agency fits this bill?

3 AUTONOMOUS AGENTS

When it comes to machines, there can be genuine disputes about whether they count as *autonomous* agents. (When people take it to be a real question whether machines are agents, perhaps they have autonomous agency in mind). We think of people, and of non-human animals, as not being autonomous agents to the extent that they're being controlled by *other* agents (human or otherwise). To the extent that I act as your slave, or handmaid

[1] Dept. of Philosophy, University of Reading, email: j.m.preston@reading.ac.uk

perhaps, my autonomy is weakened, diminished, I'm no longer 'my own man'.

But the fact that something is not in the control of any other agent can't be the only reason we think of it as an autonomous agent. A machine which is 'out of control' is not *ipso facto* autonomous, or at least not *ipso facto* an autonomous *agent*. And chemical agents can do their thing without ever being under human control, but that doesn't make them *autonomous* agents. (It seems more correct to say that such stuffs are neither autonomous nor heteronomous).

Autonomous agency is more a matter of controlling one's *own* actions. (This is what chemical agents can't do). When it comes to very sophisticated machines, like robots, we do (or at least can) distinguish between those that have more autonomy and those that have less, or none at all. And roboticists make that distinction largely in terms of the extent to which these devices control their own operations. (I don't want to presuppose that operations = actions).

To what extent, then, can machines be thought of as controlling their own operations? I think the answer is: to a limited but *increasing* extent. That is, I suggest that we're *more and more* likely to think of machines, especially robots, as controlling their own operations. Robots are always agents, and ever more sophisticated ones will be ever more autonomous. Thus, even if the notion of an autonomous agent *per se* is a notion with a kind of built-in threshold, they will at some point count as autonomous agents. Even nowadays, there's nothing wrong with *calling* certain robots agents 'autonomous', since this need do no more than signal the fact that they're not (or not continuously) under human control [see 4]. We all grasp what calling a vehicle 'autonomous' might mean – it means that it's not being fully controlled by any human agent (whether their designer, maker, vendor, or passenger).

This means that the ideas of self-control and autonomy are parts of what Daniel Dennett calls the 'intentional stance' [5] that we apply more and more readily to things other than animate beings. But autonomy and self-control will not be the key to *moral* agency. They are relatively 'thin' notions, and while they may be necessary conditions for moral agency, they're nothing like *sufficient* conditions.

4 INTENTIONAL AGENCY?

Another question, and perhaps what some will think of as the important question, is whether robots and machines are, or can be, *intentional* agents.

For that, the bar is significantly higher. To be an intentional agent is to be *the kind of agent that can have intentions*. I think this is obscured by the fact that the acts and activities of non-human agents, and perhaps even robot agents, can legitimately be characterised to a certain extent in intentional terms, i.e., in terms of beliefs and goals.

But this is because 'intentional terms' don't necessarily involve *intentions*. Dennett's important and influential idea of 'the intentional stance', which embodies the idea of 'intentional terms', muddies the waters here (deliberately or otherwise). We do apply to machine agents certain aspects of this 'stance', as he conceives it, but *not* its genuinely *intentional* aspects. Maybe there's some confusion caused by the idea of 'intentional *verbs*' here? The verb to believe is (as we know from Chisholm, Quine, etc.) an 'intensional verb' (a verb with certain logico-linguistic features), but that doesn't make it an *intentional* verb. Genuinely intentional verbs are those which pick out activities which make sense only on the supposition that the agent has intentions. But believing that *p*, although it may typically interlock with intending something or other, isn't itself a matter of intending.

Even if we allow that the acts and activities of machines can be characterised in 'intentional terms', this isn't to say that their acts and activities are *intentional*. They're not, since the agents in question can't literally be credited with intentions. We know that their operations can be understood, fully, in terms that don't mention intentions, because we know that they can be characterised in purely *mechanical* terms (as a subset of 'mechanical terms' I include probabilistic terms here). In this respect, I side with John Searle, whose claim is that machines have 'derived' intentionality, but not 'intrinsic' intentionality. It's not merely it *being* the case that the operations of machines can be understood in mechanical terms, but *our knowing that it is the case*, which makes them fail the grade. That's what gives them the specific place they occupy in our conceptual scheme.

So if, as some suppose, we could some day understand *human* actions in purely mechanistic terms, our conceptual scheme would have to change, *altering* our conception of human beings (and perhaps our conception of machines, too), and siting those new conceptions closer together.

5 THE INTENTIONAL STANCE

However, Dennett is quite right that unless we understand the inner workings of computational devices (and the connection of those workings with their environments), we're pretty much forced to take what he calls 'the intentional stance' towards them in order to understand, explain and (if we're lucky) predict their behaviour.

There is, of course, when it comes to artefacts, what I would call a *non-serious* use of intentional or mentalistic terms. All of understand what's meant when, waiting for any kind of device to respond, one says 'it's thinking about it', or 'it thinks I haven't paid yet'. The device in question may be a complicated one whose operations involve computation (like a computer, or a photocopier), but it could also be as simple as a petrol-pump, a ticket-machine, or a washing-machine.

Even at the very lowest end of this scale, it's possible to find some people who would not think of such uses as non-serious. John McCarthy, for example, famously said he thought that thermostats have beliefs (like 'it's too hot in here', 'it's too cold in here', etc.). (He wasn't talking about modern thermostats, which involve computation, but very simple mechanical thermostats, back in the 1970s). To challenge that I guess one would have to argue that having beliefs involves having discriminative capacities that interlock with one another, that beliefs have a more sophisticated manifestation than merely a *single* response, that beliefs don't *have to* result in action, that for a thing to believe it must be capable of having beliefs about a *range* of things, etc. Beliefs also have to interlock with desires that have these same features, and are capable of varying in intensity, too.

The applications to artefacts of intentional terms that I have in mind, though, go beyond this kind of non-serious use. Because most people have no programming skills and very little understanding of how computational devices operate, we have 'taken over' a certain range or kind of vocabulary which allows

everyone to communicate about very basic aspects of their functioning (and, equally commonly, malfunctioning). It should be no surprise that, despite computers being something genuinely new, an *invention*, this way of talking is no invention (like a genuinely new language), but rather a subset of an existing way of talking which is familiar to all of us, a subset of just that way of talking which Dennett calls 'the intentional stance'. (And perhaps this move is nourished by what some consider to be a universal human tendency to anthropomorphise?).

That is, all of us, even if we *resist* the tendency, understand what is meant by talking about computational devices in intentional terms. We talk about such devices in this way not only when they malfunction, or refuse to co-operate, but also when they operate according to plan.

6 GOALS, MOTIVES & INTENTIONS

As yet, we only apply certain *aspects* or parts of the intentional stance to machines, or to robots. The part we are *most* prone to apply, as far as I can see, is the basic apparatus of explaining the actions of a device in terms of its *goals* and its '*beliefs'* (or knowledge). All of us talk of them as 'searching for' information, and finding it, of having goals, and attaining them. If the device is sophisticated enough, *unlike* simple thermostats, we can also think of its operation in terms of its *strategies*. And those of us who know a bit more about AI and the programs involved have no trouble talking of them in even more sophisticated intentional terms.

We don't yet apply to machines those aspects of the intentional stance which attribute to agents *motives* or *intentions*. But goal-seeking behaviour can look like motivated or intentional behaviour, and so sometimes we can be confused about this.

7 MORAL AGENCY & MORAL THINKING

Genuine moral agency requires genuine intentionality *and* thinking. Why? Because moral agency involves the capacity to do things in the light of one's considering them good, or right, or obligatory, etc. This means that there's no reason to think that we are (yet) developing genuinely moral machine or robot agents.

Certainly robots, such as self-driving cars, and military drones will, when properly programmed, *not* do certain things, like drive or fire into a crowd of civilian pedestrians. But, unlike you and me, they won't fail to do these things *because they know they're wrong*. They will fail to do them because of their programming.

I don't mean to imply that *everything* that robots do is done 'because they've been programmed to do it'. I'm sure that's not right. But when it comes to life or death decisions we *are* going to want to retain control over robots, and we are going to want to ensure that their not doing things like this, not killing people in these ways, is as close to hard-wired as we can get. It's this *specific* feature of decisions about the most morally weighty issues which means we won't rightly think of the machines in question as having done the moral thinking involved. After all, if we *did* think of them as having done the moral thinking, we'd have to allow that they might come to a *different* decision as the result of that thinking, a different decision that would result in actions unacceptable to us.

8 MORAL AGENCY & MORAL PATIENCY

When it comes to morality, it's not just agency of a certain kind that matters. Patiency of a certain kind matters, too. That is, it makes a difference whether the being in question can be said not only to be harmed, but to *suffer*. Without some capacity for suffering, the notion of an agent being reprimanded, rebuked, or even *punished* for what it has done cannot get a grip. Animate agents are in a good position to be credited with the capacity for suffering. But what about machines?

Machines can certainly be *harmed*, but then so can paintings. We don't yet think of machines as capable of the relevant kind of *suffering*, though. Suffering is tied too closely to the biological to be credited to agents whose constitution isn't biological. No matter how much the robot arm squirms (as it were) when a heavily-loaded pallet falls on top of it, the robot is 'suffering' only in the sense in which one's car might suffer from being vandalised, or one's savings might suffer during a period of inflation.

I don't want to claim that the capacity for suffering is an absolute prerequisite for being counted as moral agent. Moral agency need not go together with moral patiency. Notably, on the one hand we count children as being moral patients (capable of the relevant kind of suffering) long before we count them as moral agents. And, on the other hand, the God of the Abrahamic faiths is often thought of as a moral agent, but one who isn't capable of suffering.

Even if the capacity for suffering isn't strictly a logical prerequisite for being counted as moral agent, though, I suspect that the extent to which artefacts will be credited with moral agency will be affected by the extent to which we think of them as capable of genuine suffering. That is, I would predict that until we are willing to credit machines and other robot agents with the capacities for thought, intention *and* suffering, we won't really think of them as moral agents.

9 MORAL RELEVANCE

A certain red herring has made an appearance in the literature here, because some people have argued that machines are *morally relevant*. People who argue thus take themselves to be arguing *against* what they call the 'neutrality thesis', according to which 'machines are neutral means for humans ends'. According to this theory, 'artefacts have no moral relevance and only human agents can be held responsible for what is done with artefacts', ([6], p.421).

We might well be suspicious when a 'theory' is formulated in this way, as a conjunction of considerations so loosely related. Let's consider it with respect to machines (not just artefacts) and take it apart into its resulting conjuncts, the idea that machines aren't morally relevant, and the idea that only humans can be held responsible for what is done with machines.

What is moral relevance supposed to be? The notion seems horribly vague. Christian Illies and Anthonie Meijers describe the views of one person who insists that artefacts are morally relevant, Peter-Paul Verbeek [see 6], as follows:

"In the Moral Relevance Debate, Verbeek seeks to eradicate the view that only the intentions of designers, producers, or users of artefacts can be evaluated in moral terms. In his opinion technological artefacts themselves are morally relevant, because of their mediating role. They affect the quality of our lives, they make us aware of morally relevant distinctions or phenomena... and they even force decisions upon us". ([7], p.424).

Apparently without defending the view that artefacts are moral agents, Verbeek does defend the idea that 'moral agency is distributed over both humans and technological artefacts' ([8, p.24]).

The view that Verbeek is opposing, that 'only the *intentions* of designers, producers, or users of artefacts can be evaluated in moral terms' is surely too limited – things other than intentions (e.g., actions, policies, strategies, etc.) can be evaluated morally. When it comes to these matters, actions and strategies involving both humans and machines can certainly be evaluated morally. And the machines in question certainly can be said to be partly responsible for the outcomes. But is this sense of responsibility anything more than the thin, causal sense, the sense in which a rock can be responsible for crushing one's car? The real question is whether anything other than the humans involved can be credited with *moral* responsibility.

Moral agency, though, is a precondition for moral responsibility. Unless something can be credited with the former, it doesn't get onto the scale of the latter. (After all (leaving aside the important case of omissions) one couldn't legitimately be held responsible for something one hadn't done). So, in my view, someone like Verbeek *will* have to commit to the idea that idea that machines can be moral agents if he wants the conclusion that they can be morally responsible.

The idea that 'only humans can be held responsible for what artefacts *do*' would need disambiguating: it's false if 'responsible' includes *any* kind of responsibility, but true if it refers only to moral responsibility. And the related idea (the second conjunct of the 'neutrality thesis', which Verbeek opposes) that 'only humans can be held responsible for *what is done with* artefacts' will survive as long as we hear 'being held responsible for' as referring to *moral* responsibility. We're not going to be able to hold military drones themselves (as opposed to their operators, etc.) morally responsible for their actions.

Arguing this, however, doesn't mean disagreeing with Verbeek on the question of the 'moral relevance' of machines, if that means what he takes it to mean, that is, that machines are 'morally relevant, because... they affect the quality of our lives, they make us aware of morally relevant distinctions or phenomena... and they even force decisions upon us'. It would be difficult to disagree with the first and third of these features, at least. So the 'neutrality thesis', formulated so as to include the denial of moral relevance, should indeed be given up.

But, whether or not this is the right way to characterise moral relevance, it does not equal moral agency. All sorts of things, other than (and sometimes far more 'basic' than) machines, are morally relevant. Landslides, earthquakes, and diseases, for example, are agents which certainly affect the quality of our lives, and force decisions upon us, so in the terms of this debate they count as 'morally relevant'. But to think of them as *moral* agents would be to indulge in superstitious thinking.

So although I'm sceptical about machines being moral agents, I'm in no way committed to the neutrality thesis. Machines *are* morally relevant, and they are so at least partly in virtue of their being *agents*. But this doesn't make them *moral* agents. One of the lessons here, though, is that the idea of 'moral relevance' is so vague as to be confusing in this debate.

10 WHAT CAN MACHINES DO?

What ranges of actions can machines, and in particular robots, be said to perform? Action-verbs of different kinds have different kinds of preconditions, and there's a kind of 'scale' involved.

Action-verbs of the most *purely* physical kind have the fewest such preconditions. In order to count as satisfying descriptions of physics, such as 'moving', 'falling', 'rotating', etc., an object just has to be a physical object. Almost anything physical can do these things, that is, perform these actions.

Action-verbs of a less purely physical kind are reserved for agents with *bodies* of certain kinds. Kicking, grasping, chewing, nodding, etc., require having the kind of body-parts with which agents perform actions of those kinds.

Action-verbs of the *mental* kind, including what philosophers call 'intentional' verbs, typically presuppose that the agent is minded. However, verbs of this kind often have what we might call an *end-state* reading in which it's not necessary that this precondition is satisfied. That is, verbs of this kind are typically such that we can think of the action as having been performed *as long as the end-state has been attained*.

This, I have argued ([9]), is the feature of such verbs that we take advantage of when it comes to digital electronic computers. We think of them as, e.g., calculating because the end-result of the process of calculation (i.e., the calculation) gets delivered, and of playing chess because chess pieces (or rather, their electronic representations) get moved across chess-boards in ways that respect the rules of chess. There should be no surprise about this: in our increasingly instrumental societies, after all, delivery of the end-result is often what *matters*. So it's now churlish to refuse to allow that computers calculate, play chess, etc.

Computers, then, genuinely do what we talk of them as doing. But 'what they do', the actions they can be credited with, are the bringings-about of the *end-results* that their operations attain, not necessarily the action(s) *we* would do (or have to do) in order to attain those same end-results. *Thinking*, however, is an aspect of the intentional stance with which we are not yet ready to credit them (*except* in the non-serious sense, but remember that non-computational devices can be described in *such* terms). So machines *can* be credited with actions, but only with actions of the non-thought-involving kind.

What range of actions does this restrict machines to? Let's think about one kind of action which might be thought of as most clearly or obviously 'morally relevant'. Military drones and driverless cars can certainly *kill* people (so can rocks, and diseases, of course). Killing someone is (to a first approximation, anyway) causing them to die, and *lots* of agents can be credited with that.

However, killing someone is one thing, *murdering* them is another. Agents won't be credited with actions picked out by verbs which inevitably imply *moral* properties (like 'to murder', 'to steal', 'to betray', 'to rape', 'to forge'), until we're willing to credit them with actions picked out by verbs which imply thought (like 'to think', or 'to deliberate', 'to ponder', etc.).

With respect to this latter kind of verb, it's true that, as machines become more and more sophisticated, we (and by 'we'

here I mean most of us, not just some computer scientists) *may* come seriously to think of them as thinking. This isn't guaranteed – it may not happen. But if it does happen this process will surely take advantage of the fact that the verb 'to think' has an end-state formulation: that is, as well as talking of what people are *thinking*, meaning what's going through their minds, we also talk of what they *think*, *simpliciter*. Belief (for this is what 'to think that so and so' means) may thus be the Trojan Horse inside which gets smuggled in the core notion of the mental act, *thinking*.

That is, even the list of core intentional verbs which we don't yet apply to machines *may* contract if and when we develop machines that we *do* want to credit with achievements of this kind. After all 'to calculate', 'to compute', 'to play chess', etc. would originally have been on the list – that is, before the age of computers, we would not originally have considered the possibility that these activities could be done by anything less than a human agent. So as our technology develops, we may begin to think of machines as doing (that is, genuinely *doing*), some of the things we now think of them as incapable of doing.

However, it won't change with respect to the first kind of verb, verbs like 'murder', 'betray', 'rape', etc., until a lot later, until we seriously think of machines as moral agents. This will require not only our thinking of them as *thinking about moral issues*, but our thinking of them as *knowing what's right and what's wrong*. And that achievement, I suggest, is still quite some way off. It may never happen.

11 CONCLUSION

Agency is much more widespread than thought. Although machines are not genuine thinkers (but merely things that can be treated as if they are thinking), machine agents *are* genuine agents. They genuinely accomplish things. But (unless and until we *do* come to think of machines as thinking, having intentions, etc.) the agency with which they can be credited is a limited subset of *human* agency, limited by their not being thinking things. And so although machines are 'morally relevant', they aren't moral agents. For that, we would have to think of them as knowing what's right and what's wrong.

REFERENCES

[1] M.Callon, 'Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St. Brieuc Bay', in J.Law (ed.), Power, Action and Belief: a New Sociology of Knowledge?, (London: Routledge, 1986), pp.196-223.

[2] J.Johnson, 'Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer', Social Problems, vol.35, 1988, pp.298-310.

[3] B.Latour, 'Where are the Missing Masses? The Sociology of a Few Mundane Artifacts', in W.E.Bijker & J.Law (eds.), Shaping Technology/Building Society: Studies in Sociotechnical Change, (Cambridge, MA: MIT Press, 1992), pp.225-258.

[4] D.G.Johnson & M.E.Noorman, 'Responsibility Practices in Robotic Warfare', Military Review, May-June 2014, pp.12-21.

[5] D.C.Dennett, The Intentional Stance, (Cambridge, MA: MIT Press, 1987).

[6] P-P.Verbeek, What Things Do, (University Park, PA: Pennsylvania State University Press, 2005).

[7] C.Illies & A.Meijers, 'Artefacts without Agency', The Monist, vol.92, 2009, pp.420-440.

[8] P-P.Verbeek, 'Obstetric Ultrasound and the Technological Mediation of Morality: A Post-Phenomenological Analysis', Human Studies, vol.31, 2008, pp.11-26.

[9] J.M.Preston, 'What are Computers (If They're Not Thinking Things)?', in S.Barry Cooper, A.Dawar & B.Löwe (eds.), How the World Computes: Turing Centenary Conference and 8th Conference on Computability in Europe, CIE 2012, Cambridge, UK, June 2012, Proceedings (Berlin & Heidelberg: Springer-Verlag, 2012), pp.609-615.

Runaway Concepts for Robotics and AI: Law, Technology and the Posthuman

Aurora Voiculescu[1]

STELARC: "Who are you?"
PROSTHETIC HEAD: "That's not a meaningful question. What is important is what happens between you and me. It's what happens in the space between us that matters. In the medium of language within which we communicate, in the culture within which we've been conditioned at this point in time in our history." (Stelarc, *Prosthetic Head: Intelligence, Awareness and Agency*, Performance/installation 2005).

Keywords: *agency; moral agency; legal agency; personhood; machine ethics; electronic personality; responsibility; legal responsibility*

This paper advances a number of essential reflections related to the complex theoretical and sociological foundations according to which the concept of agency is deployed. It looks, in particular, at the intersection of legal theory and technology and, more specifically, at the intersection of legal theory and AI. Whether one believes or not, trusts or not, all the promises put forward in relation to the AI technology, it is by now undeniable that the concept of agency, with the added notions of autonomy, rationality, action, intention, responsibility, will be considerably challenged in the coming years. For the time being, these concepts are lying in wait, in need of theoretical explanation and conceptual foundation and refinement. They feed into ethical and legal normative discourses, advancing principles and standards that build on loose premises and foggy concepts.[1] The paper begins its argument from an idea exemplified in the legal discourse as social practice, namely that agency is associated with entities – such as the human – that are often in fact pseudo-unities proposed as facts.[2] These pseudo-unities set up oppositions[3] – such as the one between the mechanical and the organic, or the one between the human and the machine[4] – that may arbitrarily separate those who are included in and those who are excluded from a shared conceptualisation or practice, such as the one of agency, autonomy and rationality.[5]

The philosophical debate on the notion of agency and its related concepts has always informed the legal discourse of responsibility, liability and legal personhood. Yet, the legal normative discourse has often appropriated these philosophical concepts and made them its' own in ways that do not always resonate with the original conceptual framework, and even less with the biological science. This process of 'transplant' of concepts is of paramount importance to addressing the social challenges that stem from the advent of AI in society. The contention of this paper is that, while the legal normative discourse brings forward an already complex normative make-up, that works with the notion of agency, this complex make-up is challenged when entering into contact and even competition with technology as social discourse. From this perspective, technology can itself be seen as a normative discourse, as yet another normative environment or as the host discourse, where the legal concept of agency and the notion of agenthood come to put new roots.

Reflecting on the bridge between the mechanical and the organic[6] and on the conceptual basis of such an encounter becomes imperative if we are to bring light to a realm that is full of conceptual pitfalls. In this sense, the paper looks, from a variety of theoretical perspectives, into the rationale of agency and responsibility and into the way in which these may inform an evolving normative discourse that can hope to address an encounter with AI technology at its highest degree of evolution, while at the same time remaining applicable to more mundane contemporary challenges. Building upon the deconstruction and analysis of the foundation of agenthood, the paper looks into the way different theories have acknowledged various forms of agency (human, animal, social, vicarious, electronic) and into the impact each of the analysed perspectives could have in the present debate on moral and legal agency on an AI platform.

Agency, together with intentionality and action, constitute the theoretical and sociological triptych that accompany the metaphysical dimensions of autonomy and rationality.[7] In the search for an ideal formula for the distribution of normative responsibility (moral and/or legal), a formula that ought eventually to have the potential to be formally sanctioned, one would have to answer the basic, but essential questions related to these three concepts in the context of AI. The discrete individual, seen as an autonomous and rational agent, is described by the normative assumptions of agency, intentionality and action both in the moral normative discourse and in the legal normative one. Yet, in law, these can easily be identified as pseudo-unities. What about other entities? What about AI artefacts? Can we conceive working with similar, homologous assumptions in defining their social parameters of action and reasons for action? Could we seriously envisage – as the European Parliament recently suggested – an 'electronic personality' that would bind autonomy and rationality in an AI agent? What would count as acts performed by this type of agent, and what should be considered as its reasons for action or as its 'intention'? Beyond the philosophical implications, a normative discourse (legal

[1] Westminster University Law and Theory Lab, email A.Voiculescu@westminster.ac.uk

or ethical) will need to 'transplant' and make those assumptions its own and answer these questions in its own language#. However, the answer to such questions has never been straightforward and the concepts of agency, autonomy, rationality have often changed substance and consistency as part of a normative discourse. Not all humans are considered as imbued with autonomy and rationality from this perspective, even less so in the past, nor have other animals always been denied such qualities.[8] Equally, the notion of collective agency – in both philosophy and law - adds to the nuances that these concepts have acquired. Drawing on a number of such points of tension and debate within the sphere of the ethical and of the legal normative discourse(s), this analysis identifies key theoretical perspectives from which the intersection between the notion of agency and AI can acquire a certain degree of consistency and solidity.

Acknowledgements

I am very grateful for the comments received from the three anonymous reviewers and to the organisers of the Symposium.

---

[1] Susan Leigh Anderson and Michael Anderson, 'A Prima Facie Duty Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a Dialogue with Ethicists' in Michael Anderson and Susan Leigh Anderson (eds), *Machine Ethics* (Cambridge Univ Press 2011)

[2] Alexis Dyschkant, 'Legal Personhood: How We Are Getting It Wrong' (2015) 2015 University of Illinois Law Review 2075; David Fagundes, 'Note, What We Talk About When We Talk About Persons: The Language of a Legal Fiction' (2001) 114 Harvard Law Review 1745.

[3] Steven S. Kapica, '"I Don't Feel Like a Copy": Posthuman Legal Personhood and *Caprica*', (2014) 23 Griffith Law Review 612.

[4] Bert-Jaap Koops, Mireille Hildebrandt and David-Olivier Jaquet-Chiffelle, 'Bridging the Accountability Gap: Rights for New Entities in the Information Society?' (2010) 11 Minnesota Journal of Law, Science & Techonology 2.

[5] Larry May, *The Morality of Groups* (Reprint edition, University of Notre Dame Press 2001)

[6] Joanna Zylinska, 'Is There Life in Cybernetics? Designing a Post-Humanist Bioethics' in Rosi Braidotti, Claire Colebrook and Patrick Hanafin (eds), *Deleuze and Law: Forensic Futures* (2009th edn, AIAA 2009); Rosi Braidotti, Claire Colebrook and Patrick Hanafin (eds), *Deleuze and Law: Forensic Futures* (2009 edition, AIAA 2009)

[7] David Copp, 'The Collective Moral Autonomy Thesis' (2007) 38 Journal of Social Philosophy 369; Kirk Ludwig, 'The Argument from Normative Autonomy for Collective Agents' (2007) 38 Journal of Social Philosophy 410.

[8] Christopher D Stone, *Should Trees Have Standing?: Law, Morality, and the Environment* (3 edition, Oxford University Press, USA 2010)

The Three Worlds of AGI

Popper's Theory of the Three Worlds Applied to Artificial General Intelligence

Marta Ziosi[1]

Abstract This Capstone applies Popper's Three-worlds paradigm to the academic discourse on Artificial General Intelligence (AGI). It intends to assess how this paradigm can be used to frame the opinions of scientists and philosophers on Artificial General Intelligence (AGI) and what it reveals about the way the topic of AGI is approached from the fields of the Sciences and the Humanities. This has been achieved by means of a Literature Review reporting the opinions of main philosophers and scientists and by analysing two main projects – project CYC and project SOAR- advanced as possible ways to achieve AGI. As a result, most academics from the field of Science seem to better fit views on AGI interpreted through the lens of Popper's World 2, the world of the mind. On the contrary, most philosophers seem to better fit views on AGI interpreted through the lens of Popper's world 3, the world of the products of the human mind such as theories, knowledge and ideas. As a suggestion, this Thesis advocates the promotion of interdisciplinarity and discussion among the different academic fields.

1 INTRODUCTION

Back in 1965 the US psychologist Herbert Simon proclaimed that machines will be capable within 20 years to do any work a man can do (Simon, 1965). Nevertheless, the present state of affairs showcases how the promise has not withheld its foretelling. Why? It is a matter of timing? Or is it an illusionary idea which can avail itself solely of these empty '20 to 30-years' futurist prognoses? Opinions largely differ and many a times collide within people from different levels of expertise and belonging to different fields of research. Different opinions can be gathered from branches of Computer Science to Philosophy, from the Cognitive Sciences to Technology and Media Studies; more generally, from the fields of the Sciences to the ones of the Humanities.

Arguably, the question ought not to be of an 'all or nothing' nature but one about the approach we as humans should take towards General Intelligence. Plainly, the past years have witnessed an incredible confluence of storage of big data, probabilistic programming and sheer increase in computing power. However, computers are still not capable of engaging in some apparently easy tasks for humans. The approach should be in thinking about robots and AGIs – Artificial Intelligent Agents - not just as a technology which engages in physical and computational work. The key relies in thinking about them indeed as physical computational entities but in relation to humans[2]. Several researchers are already engaging with such an approach. The main questions which are being asked are of the kind, *'How can we and What does it mean to create an AGI which thinks?'* or '*What does it mean to create an AGI with a common sense of human society, knowledge and culture?'*.

I hypothesize that while researchers in the field of science tend to work on the first question, the ones in the humanities tend to focus on the latter. However, any potential answer to both questions fundamentally requires both computational capabilities or understanding of algorithms from the sciences and critical thinking or the heuristics of the humanities. Thus, if the intent is to reach a *generally* intelligent agent, the efforts ought to hail from an as encompassing as possible *interdisciplinary* background. To achieve that, we ought to agree on the question to ask. This is essential in order to avoid the carry-out of *miscommunication* under the illusion of *disagreement*.

This research will thus propose a framework to swiftly cut through the two different approaches in order to identify their differences in topic and purpose. The core-framework will be provided by Popper's theory of the Three Worlds. The question which instructs this Capstone is '*How can Popper's Three-worlds paradigm be applied to frame the opinions of scientists and philosophers on Artificial General Intelligence (AGI) and what does it reveal about the way the topic of AGI is approached?'*. The core information on the topic of AGI will be proffered by means of exposing the main ideas and opinions over AGI of mainly scientists and philosophers. Conceivably, a thorough analysis of these will be conducted by applying the chosen framework. The last word is left to the conclusion where a suggestion on how to deal with discrepancies in opinions will be advanced.

Finally, it is important to state that this thesis does not aim at predicting future scenarios and it aligns itself with Popper's claim that predicting technological innovation is impossible (Popper, 1979). Indeed, if humans could, they would already know how to implement it, thus leaving no logical space between the prediction and the realization of the technology. The intended relevance of this thesis is principally to provide a broader outlook on matters of AGI and it aims at breaching through the AGI discourse by Popper's toolbox of World 2 and World 3 in order to expose a potential thought-gap or discrepancy of opinions between two chief-fields. A suggestion in favour of interdisciplinarity will be advanced at the end.

[1] London School of Economics, email: M.Ziosi@lse.ac.uk.

[2] I am aware and I will not deny the importance of the physical part of the process of computation. This sentence is solely aimed at emphasising the importance of thinking about this 'physical part' in relation to human capabilities, given that the goal is Artificial General Intelligence.

2 DEFINITIONS

2a. Intelligence

To begin with, it is important to define the term 'intelligence' in the way in which it will be used in the paper. Intelligence is the '*computational* part of the ability to achieve goals in the world' (Stanford, 2017). There are varying kinds and degrees of intelligence which occur in people, in many animals and some machines. As it has not yet been decided which computational procedures ought to be called 'intelligent', it is also extremely difficult to frame a solid definition of intelligence which detaches itself from any reference to human intelligence as that is the only example at present. Thus, this definition ought not to be dogmatic throughout the Thesis but it mostly serves as a guideline.

2b. Artificial intelligence

Artificial Intelligence (AI) is '*the science and engineering of making intelligent machines'* (Stanford, 2017). AI does not necessarily limit itself to biologically observable methods. Indeed, even though brain emulation[3] is an example of AI, there are several other approaches to AI such as ones working through probability or brute force algorithms (Goertzel, 2007).

2c. General intelligence

General Intelligence is the ability to achieve complex goals in complex environments (Goertzel, 2007). The plurality of the words 'goals' and 'environments' is crucial to explain how a single goal or a single environment would not account for the word 'general'. Indeed, a chess-playing program is not to be considered 'generally' intelligent as it can only carry-out one specific task. An agent possessing artificial intelligence ought to have the ability to carry-out a variety of tasks in diverse contexts, generalize from these contexts and to construct an understanding of itself and the world which is independent of context and specific tasks.

2d. Artificial general intelligence (AGI)

A complete appreciation of the challenges encountered by the idea of 'general intelligence' in the field of AI requires a wide range of perspectives to be adopted. Correspondently, Artificial General Intelligence (AGI) is a highly interdisciplinary field. As it follows from the definition of AI, it could be said that AGI is 'the science and engineering of making *generally* intelligent machines'. As it follows from the definition of General Intelligence, AGIs are expected to solve a wide range of complex problems in several contexts. Additionally, they learn to solve problems whose solution was not presented to them as the stage of their creation. Currently, there are no existing examples of AGIs in the real world.

[3] The process of copying the brain of an individual, scanning its structure in nanoscopic detail, replicating its physical behaviour in an artificial substrate, and embodying the result in a humanoid form (aeon).

3. STATE OF AFFAIRS IN AI

The present section will acquaint the reader with a brief background on the history of AI and AGI (first sub-section), the approaches to AGI (second sub-section) and finally, projects and possible solutions (third sub-section).

3a. A bit of history of AI and AGI

In 1956, after the first programmable computer was invented, the genesis of a new field called 'Artificial Intelligence' was announced at a conference at Dartmouth College in New Hampshire (Brey, 2001). This field had the ambition to supply computers – by means of programming - with some sort of *intelligence*. Even before that, the scientist Vannevar Bush had already proposed a system which had the aim to amplify people's own knowledge and understanding (Bush, 1945). It was only five years later when the now celebrated Alan Turing wrote a paper centred around the idea of machines being able to simulate human beings and to carry out intelligent tasks, such as the playing of chess (Turing, 1950). As such, the idea of a machine which could encapsulate some sort of conception of intelligence can already find its space in that years.

3b. Current approaches

Nowadays, there are two main views held in regard to algorithms. These two shape the different directions taken by approaches to AI. One is held by the proponents of *strong AI* and one by the ones of *weak AI*. The ones defending the former argue that an algorithm is a universal concept which is applicable to anything that works mechanically and thus, also the brain. They argue that human intelligence works through algorithmic processes just like computers. However, as the algorithmic processes regulating the brain are highly sophisticated, they do contend that there does not yet exist any man-made system comparable to it. Yet, it is only a matter of time. On the contrary, proponents of weak AI maintain that even though aspects of human thinking are algorithmic, there are critical aspects about the way humans are given to experience the world which do not fit the algorithmic picture and probably never will. Humans experience the world from sensations. These two characteristic approaches to AI also shape any groundwork on AGI. Hence, they ought to be kept in mind throughout the Thesis to better grasp the subject matter.

3c. Projects

Apart from these two main approaches, there are several projects which have been attempted through the years and which are important to present in order to better understand the nature of the concerns and points advanced in the literature review. Two projects will hereby be presented. It is important to state that they differ in approach. These two projects are the CYC project and the SOAR project.

In the mid 80s, the CYC project began as an attempt to encode common-sense knowledge in first-order predicate logic (Goertzel, 2007). The encoding process was a large effort and it produced a useful knowledge database and a specialised and

complex inference engine[4]. However, until today CYC does not *'solve problems whose solution was not presented to them at the stage of their creation*' (see AGI definition). Plainly, it does not come up with its own solutions; which is a defining feature of AGIs. CYC researchers have encoded in the system common-sense knowledge. However, this knowledge-filled database has resulted in an open-ended collection of data more than dynamic knowledge. By making use of declarative language by means of Lisp syntax[5], CYC features the ability to deduce concepts. However, differently from neural networks techniques, it still relies on humans inputting an 'unending' amount of data before outputting any result. This is one of the main critiques adduced to the CYC case. CYC enthusiasts have rushed in its defence by stating that CYC has the potential to be imported in future AI projects (Goertzel, 2007).

Adopting an opposite approach, Allen Newell's SOAR project is a problem-solving tool which is based on logic-style knowledge representation and mental activity figured as 'problem solving' expressed by a series of heuristics (Goertzel, 2007). The core of the effort behind the SOAR project is to investigate the architecture which underlies intelligent behaviour (Rosenbloom, Laird & Newell., 1993) and what constitutes intelligent action rather than knowledge. SOAR can be described as a sequence of three cognitive levels; the memory level, the decision level and the goal level. These are merely descriptive terms which are used to refer to the mechanism constitutive of the SOAR architecture (Rosenbloom, Laird & Newell, 1991). Even though it represents a great step in the AGI field, up until now the system is still a disembodied problem-solving tool lacking the autonomy and self-understanding which are expected in an AGI.

4. LITERATURE REVIEW

Notwithstanding the various pursuits for AGI implementation, the discipline was propelled chiefly from an idea. The present section will focus on the intellectual life and discourse surrounding AGI. This section lays the groundwork for the future analysis.

4a. Different worlds

Through the following paragraphs, it is more specifically presented how, through the years, the expectations and what are considered the key factors on the way to AGI have differently developed on the side of the Humanities and on the side of the Sciences. The following paragraphs ought to elucidate this claim. Even though with a risk of redundancy, it is important to state that all the scholars and great minds presented in the paragraph 'The response of the science world' come mostly from a scientific background, while the ones in 'The response of the Humanities' come mostly from a philosophy background. Some of them have also expertise in both fields. In that case, they are found in the section for which their background is stronger. The following paragraphs provide the content which will be subject to the application of the Theoretical Framework later in the paper.

4b. The stance of the science world

Influenced by the groundwork of Alan Turing, the 70s featured the creation of Putnam's 'mentalist project' (Dreyfus & Haugeland, 1974). The mentalist project was an endeavour to represent the rules that govern human behaviour and the mind by a Turing machine table that relates input and output states. Concurrently, the scientists Newell, Shaw and Simon who were in the 1950s considered the pioneers of Cognitive Simulation, announced that *'within ten years most theories in psychology will take the form of computer programs'* (Simon & Newell, 1957, p.8). George Miller himself, a distinguished psychologist at Harvard, asserted that the current developments in the study of man's understanding could be viewed as a system of *information processing* (Miller, Galanter & Pribram, 1960, p.57). The configuration of mental processes as computations was taken beyond a mere analogy.

A more critical stance towards the ability of re-creating certain mind-phenomena such as consciousness through algorithms is provided by the scientist Roger Penrose in his book, 'The Emperor's new Mind' (1989). On one hand, he claims that the mind understood as 'consciousness' cannot be computed. However, he contends that this is impossible only as long as the model is based on the idea of a Turing Machine, as the latter only *mimics* mental processes and does not progress towards any kind of 'understanding' for the machine. Even though rejecting the Turing Machine's paradigm, as many other scientists he strongly defends that more generally mental activity is *'the carrying out of some well-defined series of operations'* (Penrose, 1989, p.17). He resorts to call these operations 'algorithms'. Penrose does convene that mental activity can be represented through algorithms. Additionally, he stresses that human mental processes result in our ability to 'understand' and that is what research ought to focus on. AGIs can improve their performance by experience through a sort of 'feedback system' for performance improvement. According to Penrose, this might account for some kind of 'understanding'.

Another scientist who widely confronted the assumptions underlying AGI implementation is Murray Shanahan[6]. Interestingly, as also Penrose proposed, he figures the main challenge on the road to AGI as a matter of endowing a system with '*common sense understanding'*. Howbeit, Shanahan considers '*common sense understanding*' to need to blend with *creativity*. He calls both these elements 'cognitive ingredients' and while describing AGI, he locates it in what he calls '*the space of possible minds'* (Shanahan, 2016). Thus, he adopts a mind-stance. In the space of possible minds, AGI can figure either by means of 'whole brain emulation'[7] or by constructing

[4] More insights can be found on the site: www.cyc.com

[5] Lisp is the second-oldest high-level programming language favoured for research in Artificial Intelligence. It allows to interchangeably manipulate source codes as a data structure. His command line is called a *Read-Eval-Print-Loop* as it *reads* the entered expression, *evaluates* them and *prints* the result (Chisnall, 2011).

[6] It is important to state that, even though Murray Shanahan is an expert in Cognitive Robotics, he also engaged in several philosophical work.

[7] The process of copying the brain of an individual, scanning its structure in nanoscopic detail, replicating its physical behaviour

an artificial brain which matches a statistical description of a new-born's central nervous system. Even when Shanahan admits that the human brain is not necessarily the starting point on the path to AGI, he proposes different architectures such as brute force search algorithms and machine learning techniques which approach the problem on terms of computation (Shanahan, 2016). Indeed, he convenes that 'human thinking' can be instated through computation, whether they resemble the brain or not.

4c. The stance of the humanities

On the side of the humanities, the philosopher Hubert Dreyfus claims that AGI is based on a boastful epistemological assumption. This assumption implies that all knowledge is formalizable. Plainly, humans' thoughts and actions have produced a body of knowledge on which human reality feeds itself and stands on. Howbeit, AGI assumes that this body of knowledge can be expressed in context-independent formal definitions and rules (Brey, 2001, p.5). He asserts that while these formal rules can successfully describe human knowledge, they cannot be used to reproduce it. In fact, the application of these rules is actually context-dependent. Hence, he contends that there is a body of knowledge - constitutive of human reality - which ought to be acknowledged in AGI implementation. However, at the same time he stresses that this knowledge is too dependent on circumstances and on context to be successfully objectively formalized; this is where the main challenge lies.

Another philosopher who adds a valuable contribution to the topic is John Searle. Dreyfus and Searle agree on the fact that (Strong) AI relies on another mistaken assumption. Strong AI figures intelligent systems as symbol processing systems (Brey, 2001, p.4; Searle, 1990, p.26). According to this view, thinking merely consists in symbol manipulation rather than meaning and human knowledge. Additionally, such an assumption furthers the idea that the mind stands to the brain as a program stands to the hardware. Searle however, strongly refutes this view. He claims that minds are not programs. In fact, programs are formal, syntactic and thus, they are sufficiently defined in terms of symbol manipulation. For example, a line of program can be 'if 01, then print 1'. In this case, a program does not need to understand or have *knowledge* of what '01' means in order to execute 'print 1' and to move from symbol '01' to symbol '1'. Differently, human minds have *mental contents* (Searle, 1990) and the linguistic understanding which happens between people who intend to share mental contents requires a semantic framework as provided by the net of human knowledge. This is what enables the conveying of meaning. As it is presently defined, strong AI appears to overlook this difference which is instead crucial when dealing with 'general intelligence'.

5 THEORETICAL FRAMEWORK

This section presents the theoretical framework which provides the lens through which the literature will subsequently be analysed.

in an artificial substrate, and embodying the result in a humanoid form (aeon).

5a. Core argument: Popper's three worlds

Karl Raimund Popper was born in Vienna in 1902. He is one of the most prominent philosophers of Science. Karl Popper is more commonly associated with Critical Rationalism and his most acclaimed work is about Falsificationism and the evolution of objective knowledge in scientific inquiry. A special focus will hereby be dedicated to his pluralist view on reality.

Popper advocates a pluralist view of human reality. According to him, there exist three 'Worlds' or 'sub-universes' (Popper, 1979). World 1 consists of physical bodies. Plainly, elements of it are physical living and non-living objects such as stars, stones, animals and plants. World 2 is the world of conscious experience. It is the mental and psychological world filled with subjective experiences, mental states like pain and pleasure, perceptions and intentions. It is what humans think about the world as they try to map, represent, hypothesize or anticipate in order to maintain their existence in an ever-changing place. Finally, world 3 is the world of the products of the human mind. This broadly includes languages, songs, paintings, mathematical constructions, theories and even culture.

Popper strongly advocates not only the existence of the products of the human mind, but also their being real rather than fictitious. As far as these have a causal effect upon us, they ought to be real. Products of the human mind, for example scientific theories, have proven to have an impact on the physical world by changing the way humans build things and utilize them. Popper believes that the causal impact of world 3 is more effective than scissors and screwdrivers (Popper, 1979). Furthermore, even though elements of World 3 are generally instantiated in a concrete object of World 1 – books, physical components of a computer… -, it is not a necessary condition that they be so expressed (Sloman, 1985).

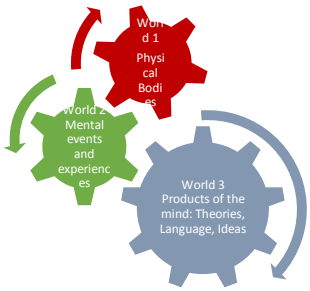

Figure 1. Popper's Three Worlds visualization

This simple above visualization suggests Popper's acknowledgement of the interaction between the three worlds. According to Popper, World 3 theories or plans always ought to be primarily understood by a mind in World 2 before they be operationalized. Withal, the theory itself and its operationalization have effects on World 1 physical objects. An example can be purported by Einstein's Theory of Relativity. The scientific community had to first subjectively grasp the content of the Theory of Relativity before this could be applied to change the physical reality. Hence, World 2 proves itself to be a necessary intermediary between World 3 and World 1. Likewise, as Einstein's special Theory of Relativity lead to the creation of the atomic bomb, World 3 impacts World 1.

Finally, both for the specific purpose of this research and to follow Popper's emphatic concern for this distinction, we ought to precisely differentiate between 'thought processes' and 'thought contents'. The former belong to World 2 while the latter to World 3 (Popper, 1979). Even though these two might appear to be interchangeable, they are fundamentally and foundationally different. It is paramount to understand that the process of thinking is unlike the knowledge which this process itself unveils. This distinction ought to be sheltered in the reader's mind as it gains momentum in the following paragraphs.

6 DISCUSSION

6a. Popper's three worlds

Programmatic processes – ex. Algorithms - and the data which they output and process act in interplay. For example, intelligent systems' internal algorithms are designed to deal with the data they are inputted with and the way they process these data consequently modifies the output. These processes – such as algorithms – and data – such as big packages of information – both ought to exist and co-exist in an AGI system and they have an impact the one on the other. While several algorithms in AGI aspire to imitate *thought processes*, the knowledge or data which they process and output can be thought of as the *content* which is the *product* of these processes. As Karl Popper stressed, *thought processes* – related to mental events and states - and *thought contents* – related to objective contents of thoughts – belong respectively to two different 'worlds' and hence, they are foundationally and fundamentally different (Popper, 1978). Indeed, the process of thinking is unlike the knowledge which this process itself unveils. Both concepts seem to unilaterally figure in the understanding and explanations of AGI, depending on from which field – Science or Humanities – the claim originates. Now, do they?

Both Penrose and Murray Shanahan build the foundations of their work on AGI on the conviction that the mind can be computed and specifically Penrose refers to AGI as a matter of 'mental processes' which manipulate information. On the other hand, philosophers such as Dreyfus claim that AGI systems ought to be deeply characterized by the character of the information which they manipulate and thus, they stress the role of World 3 *thought contents*. Popper's pluralist view helps to shed light on this subtle and yet fundamental distinction which appears to delimitate mainly the views of researchers in Philosophy and Scientists on the topic of AGI.

Arguably, if we read the topic of AGI under a World 2 lens, both subjective experience and mental tasks are key words (Popper, 1983). As per subjective experience, in the section 'Experience as Method' Popper addresses subjective empirical experience as the structured, logical description of only one world – the 'world of our experience' (Popper, 1983) - out of an infinite number of logically possible worlds. In the AGI case and for computers, the expression of their only 'world of experience' happens through binary logic and their 'mental tasks' are carried out through algorithms. Computer scientists and AI researchers adopt binary logic as their main tool and psychologists and neuroscientists primarily study mental tasks and subjective experience. Could this favour a reading of AI from a World 2 perspective?

On the other hand, in 'Epistemology without a knowing subject', Popper considers World 3's objective knowledge such as theories and ideas as something which does not need a knowing subject; as an entity independent of anybody's disposition or belief towards knowledge (Popper, 1972). Once we apply this to the context of AGI, Dreyfus would agree in the sense that we, as humans, rely on a body of knowledge that we have produced. That knowledge can be used to *describe* human behaviour. He claims that there is a body of knowledge that ought to be recognized in the implementation of AGI. Nevertheless, this last of Popper's formulations dissents with Dreyfus acknowledgement of the importance of *context* in matters of human knowledge. Indeed, Dreyfus contends that human knowledge is highly dependent of context and circumstances and henceforth, not independent of a subject. Searle would also recognize the importance of a net of human knowledge from which to derive meaning. Nevertheless, he would also disagree in the sense that for him this knowledge is highly dependent of people's dispositions towards it. Thus, even though both philosophers would stress the importance of 'knowledge', Popper's world 3 does exhaust what is important in their views.

6b. In the real world

The attempt to interpret the AGI discourse by means of the tension between World 2 and World 3, might advance a hypothesis on a possible reason why projects such as CYC and SOAR have not resulted to be successful (from section 'Projects and Possible Solutions'). On one hand, the CYC was started with the aim to encode all common knowledge. However, as it is a knowledge-filled database, it has resulted in the accumulation of data. On the other hand, the SOAR project was started with the aim to instantiate mental activity. However, as it reproduces 'intelligent action' by algorithms rather than knowledge, it has resulted in a disembodied problem-solving tool. It is clear how 'General Intelligence' cannot be reached unilaterally. While the endeavours of the CYC project might be better represented by World 3, SOAR's endeavours might be better represented by World 2. It ought to be acknowledged that in reality these two Worlds interact. Thus, it might be fruitful to think about a General Intelligent machine as something which can integrate both *though processes* and *thought contents*, the content of a theory and the subjective processing of it.

7 LIMITATIONS

One of Popper's admirable recommendations is that one ought to expose potential weaknesses of one's theories (Popper, 1983). As per this thesis, there are several factors which ought to be taken into consideration while reading it and of which the reader should be made aware of. The first point concerns the Theoretical Framework. Indeed, the backbone of the argument which derives its structure from Popper's Three Worlds cannot be said to uniformly apply to every case of the AGI discourse or research. While the framework has proven to be arguably sound for some limited cases in Science and Humanities, the panorama can supposably vary for interdisciplinary cases. Some mathematicians are also trained philosophers and vice-versa. Further research could venture in examining such cases.

Moreover, the distinction which Popper meant to draw between the Worlds appears to be an ontological one. In his 'Objective Knowledge' he presents the idea of three different ontological worlds (1972). Furthermore, in his 'Knowledge without a knowing Subject' (1972) and in his 'Three Worlds' (1979) he repeatedly stresses the existence of World 3 independently on a subject perceiving it and it justifies its existence by means of the causal impact it has on other Worlds. Given these considerations, it ought to be stressed that this Thesis utilizes Popper's distinction to try to group different *readings* or different *standpoints* on the matter of AGI. However, it does not claim any ontological difference between the three Worlds.

8 CONCLUSION

The present thesis has traversed the topic of AGI by first providing a brief account of its history, different approaches and projects. A more in-depth prospect on the matter has been presented by the Literature Review. The Theoretical Framework has served as a toolbox to analyse the AGI discourse from famous academics and scholars. At the very incipit the question was, '*How can Popper's Three-worlds paradigm be applied to frame the opinions of scientists and philosophers on Artificial General Intelligence (AGI) and what does it reveal about the way the topic of AGI is approached?'*. By the end of this Thesis, it can be argued that World 2 and World 3 can be utilized in framing and grouping the opinions of the two disciplines on the topic of AGI. More broadly, this can be framed in terms of approaching the topic by means of *thought processes* (World 2) and *thought contents* (World 3). This analysis can hypothesize discrepancies between the two 'worlds' of Philosophy and Science, when they tend to more strongly approach AGI from just one of the stances. Even though each stance provides a 'safe place' for each field, on one hand it is difficult to rely on World 3's objective knowledge and theories without taking into consideration the *mental processes* which output this knowledge. On the other, it is difficult to claim that 'understanding' automatically arises from World 2's mental processes by overlooking World 3.

In conclusion, what could we learn or advance from this analysis? Overall, the possible suggestions are innumerable but I believe that the incentivizing of interdisciplinarity can favour the opening of worldviews, communication between and within fields and finally, place the AGI discourse in Popper's World 3 where, either as a theory or as a mere human idea, it can be subject to critique. I contend that an interdisciplinary approach ought to be more cherished as it promises more realistically nuanced outcomes than trying to figure out and picture every possible future AGI scenario from each discipline. Furthermore, it can integrate the different stances from each field, transforming an obstacle into an asset. Researchers, professors but also students ought to be acquainted through their path of study with what other fields have to say and with their now still 'alien' worldviews. The ways to push interdisciplinarity on the agenda are innumerable, from curricula in schools and universities to open conferences, journals and more accessible popular events. As it is, AGI is an interdisciplinary matter in itself and it has the potential to lure people towards its topic from several angles. This would also avoid the spreading of *fear* towards the future of AI and AGI, a fear which many a times derives from miscommunication and misunderstanding. We should better concentrate together on what *is* possible rather than on what *might* happen.

REFERENCES

Basic Questions. (2017). *Www-formal.stanford.edu*. Retrieved 16 June 2017, from http://www-formal.stanford.edu/jmc/whatisai/node1.html

Hobbes, T. (1958). *Leviathan* (5th ed., p. 45). Library of Liberal Arts.

Brey, P. (2006). Evaluating the social and cultural implications of the internet. *ACM SIGCAS Computers and Society*, *36*(3), 41-48.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, *4*(3), 185-211.

Chisnall, D (2011). *Influential Programming Languages, Part 4: Lisp*.

Clark, A., & Chalmers, D. (1998). The extended mind. *analysis*, 7-19.

Goertzel, B. (2007). *Artificial general intelligence* (Vol. 2). C. Pennachin (Ed.). New York: Springer.

Dennett, D. C. (2008). *Kinds of minds: Toward an understanding of consciousness*. Basic Books.

Dreyfus, H. L. (1972). What computers can't do. MIT press.

Dreyfus, H., & Haugeland, J. (1974). The computer as a mistaken model of the mind. In *Philosophy of Psychology* (pp. 247-258). Palgrave Macmillan UK.

Dreyfus, H. L., Dreyfus, S. E., & Zadeh, L. A. (1987). Mind over machine: The power of human intuition and expertise in the era of the computer. *IEEE Expert*, *2*(2), 110-111.

Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason*. MIT press.

Frowen, S. F. (Ed.). (2016). *Hayek: economist and social philosopher: a critical retrospect*. Springer.

Goertzel, B. (2007). *Artificial general intelligence* (Vol. 2). C. Pennachin (Ed.). New York: Springer.

Hayek, F. A. (1945). The use of knowledge in society. *The American economic review*, 519-530.

Hecht-Nielsen, R. (1988). Theory of the backpropagation neural network. *Neural Networks*, *1*(Supplement-1), 445-448.

High, R. (2012). The era of cognitive systems: An inside look at ibm watson and how it works. *IBM Corporation, Redbooks*.

Hobbes, T. (1958). *Leviathan* (5th ed., p. 45). Library of Liberal Arts.

Horsman, C., Stepney, S., Wagner, R. C., & Kendon, V. (2014). When does a physical system compute?. In *Proc. R. Soc. A* (Vol. 470, No. 2169, p. 20140182). The Royal Society.

Horwitz, S. (1998). Review of Frowen, Stephen F., ed., Hayek: Economist and Social Philosopher: A Critical Retrospect.

Horvitz, E. J., Breese, J. S., & Henrion, M. (1988). Decision theory in expert systems and artificial intelligence. *International journal of approximate reasoning*, *2*(3), 247-302.

Lloyd, S. (2004) *Programming the universe*. New York, NY: Alfred A. Knopf.

Miller, G. A., Galanter, E., & Pribram, K. H. (1986). *Plans and the structure of behavior*. Adams Bannister Cox.

Penrose R. (1989) *The emperor's new mind*. Oxford, UK: Oxford University Press.

Popper, K. R. (1972). Objective knowledge: An evolutionary approach.

Popper, K. R. (1978). Three worlds. The Tanner Lecture on Human Values. The University of Michigan. *Ann Arbor*.

Popper, K. (1983). *The logic of scientific discovery*. Routledge.

Rosenbloom, P. S., Laird, J. E., Newell, A., & McCarl, R. (1991). A preliminary analysis of the Soar architecture as a basis for general intelligence. *Artificial Intelligence*, *47*(1-3), 289-325.

Rosenbloom, P. S., Laird, J., & Newell, A. (Eds.). (1993). The SOAR papers: Research on integrated intelligence.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(03), 417-424.

Searle, J. R. (1990). Is the brain's mind a computer program. *Scientific American*, *262*(1), 26-31.

Searle, J. R. (1992). *The rediscovery of the mind*. MIT press.

Shanahan, M. (2012). Satori before singularity. *Journal of Consciousness Studies*, *19*(7-8), 87-102.

Shanahan, M. (2015). *The technological singularity*. MIT Press.

Shanahan, M. (2016). *Beyond humans, what other kinds of minds might be out there? – Murray Shanahan | Aeon Essays*. (2016). *Aeon*. Retrieved 16 June 2017, from https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there

Simon, H. A., & Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations research*, *6*(1), 1-10.

Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London mathematical society*, *2*(1), 230-265.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433-460.

Bush, V. (1945). As we may think. *The atlantic monthly*, 176(1), 101-108.

Sloman, A. (1985). A Suggestion About Popper's Three Worlds in the Light of Artificial Intelligence. *ETC: A Review of General Semantics*, 310-316.

The ethics of developing AI: Why advanced AI has moral status

Irina Zudina[1]

Abstract. I argue for the moral status of artificial entities to prevent mistreatment and suffering, considering capacities to perceive experiences such as pain and pleasure. This argument is sentience-based in terms of phenomenal consciousness possibly arising from a certain degree of functional equivalence.

1 INTRODUCTION

Due to current debates about the development of High Level Machine Intelligence (HLMI), and machine consciousness undertaking tasks in several domains of everyday life, it is indispensable to discuss the moral status of such entities in our society and consider accompanying duties. Failing to do so could result in mistreating of sentient artificial minds, and consequently immense suffering [1].

Reasoning moral status based on a sentience-based criterion leads me to analyse artificial entities capacities to perceive experiences and stimuli. Further, I will briefly mention the role of phenomenality, suggesting ways to bypass the explanatory gap.

2 MORAL STATUS

Moral status is divided into moral agency and moral patienthood. While moral agency focuses on the term of sapience, in this paper, moral patienthood is justified by sentience.

With the aim of increasing pleasure and decreasing overall suffering, utilitarian philosophers justify moral status by the lowest common denominator: sentience and the ability to suffer. My argument is based on Peter Singer's / Jeremy Bentham's claim of moral status as a moral patient is based on sentience ("The question is not, Can they reason? nor Can they talk? but, Can they suffer?"[2]). This gives entities, with the capacity for phenomenal conscious experiences, such as pain and pleasure, the status of a moral patient being morally considerable [2, 3]. Such capacity for suffering and experiencing happiness, also called psychological capacities, is fundamental to have interests at all. As I will not further discuss the content of my premises in this paper, let's assume the following:

Theorem 1 *Sentient entities are moral patients and deserve moral consideration.*

Theorem 2 *Sentience is based on (phenomenal) consciousness and experiences.*

I wish to distance myself from requirements concerning the way and the level a moral patient - in this case an artificial entity - must be conscious. As consciousness is too complex to be explained at current stages of research, we can only claim to have phenomenal experiences ourselves. Distinctions between humans, different sorts of non-human animals, and artificial beings are made without having enough knowledge about consciousness in general.

This raises the issue to what extent a machine or developed AI can gain and possess these features.

3 ARTIFICIAL ENTITIES

David Chalmers's philosophical analysis on the thought experiment of mind uploading lets us make a first hypothetical step towards imagining an artificial entity to be conscious [4].

In this hypothetical scenario, gradual uploading uses very small nanotechnology devices which are inserted into the brain to "learn" the behaviour of the original biological neurons. As these devices gain enough knowledge and skills to emulate the original neuron, it replaces the original. After a time all biological components and neurons are destroyed by the replacement. It concludes a preservation of consciousness. As this example can not be applied to the "real" / current development of artificial intelligence and machine consciousness it further serves to defend a functionalist perspective and to argue against speciesism (and discrimination) towards artificial entities.

While the biological theory implies that consciousness is always based on a biological system, and a non-biological system therefore can not be conscious, the general functionalist theories focus is on the causal structure and role. According to the functionalist theory it does not matter what the being we consider to have consciousness is made of.

Therefore, agreeing with Chalmers's line of reasoning, I strongly emphasise the causal structure and de-emphasise its biological or physical composition.

Two principles out of the "Ethical Principles in the Creation of Artificial Minds" by Nick Bostrom in 2001, give further support by arguing against speciesism towards artificial entities [5].

The "Principle of Substrate Non-Discrimination" values the significance of functional equivalence over the substrate an entity is made of. In addition, the "Principle of Ontogeny Non-Discrimination" asserts the irrelevance of how an entity is brought into existence.

[1] University of Osnabrück, Germany, email: izudina@uni-osnabrueck.de

[2] J. Bentham (1789). Introduction to the Principles of Morals and Legislation, chapter 17.

Therefore, as long as there are no fundamental differences in functionality and causality between existing sentient beings and any other (artificial) beings, these requirements are sufficient for attributing moral patienthood.

As the current development of high-level-machine-intelligence and machine consciousness lacks such functional equivalence, my argument leads to the question of how fine-grained this equivalence has to be, or whether it has to be present at all, to conclude moral status. As there are things with little functional equivalence, which we do not define as conscious, it is necessary to draw a line or consider other suggestions, like the argument of mere machines' teleological interests, by John Basl [6].
But as my line of reasoning is based on sentience and the capacity to perceive pain and pleasure, I will further face the role of phenomenality and qualia.

4 THE ROLE OF PHENOMENALITY AND QUALIA

I claim that phenomenality and qualia are essential for conscious experiences leading to sentience and moral status.

But due to the lack of understanding concerning phenomenal consciousness and the explanation of qualia [7], picking up the previously used functionalist approach to explain the perception of pain and pleasure seems plausible. This leads me to argue that (not precisely fine-grained) functional equivalent input- and output-devices making the perception of stimuli possible, may give a base for the perception of phenomenal experiences and qualia as well as good reason to believe an entity to have phenomenal consciousness. Still, it is necessary to bear in mind that newly developed artificial entities might differ from anything we know of, concerning input-functionality and psychological perceptions [8].

This uncertainty further leads me to contemplate different approaches used to prevent mistreatment and make artificial entities morally considerable.

5 ALTERNATIVE APPROACHES AND CONCLUSION

As mentioned before, John Basl suggests an approach to regard artificial beings interests [6]. He argues that at the current state of development and research we can not prove psychological states or psychological interests in machines, but can consider machines specific teleological interests which are not in need of any mental life or states.
Alternatively, in case we can not prove the functional equivalence or the consciousness of artificial beings, I plead for the application of the principle of prudence. This makes us treat such entity as having a moral status to "play safe" and prevent crucial repercussions.
Apart from the principle of prudence I add the utilitarian idea of "expected value", which is estimated by calculating scope and probability.
The latter two are also further incentives to dedicate oneself to the research done on the ethics of artificial intelligence.

I conclude moral patienthood for advanced artificial entities, in particular High-Level-Machine-Intelligence, with i) (not precisely fine-grained) functional equivalence to existing moral patients with ii) present input devices to perceive environmental stimuli. In other cases, the application of i) teleological interests' consideration, ii) principle of prudence, and iii) "expected value" seem plausible approaches to prevent mistreatment and suffering, as long as we do not have deepened knowledge about qualia and artificial entities' phenomenality.

Acknowledgements

I would like to thank my graduation supervisors Uwe Meyer and Sebastian Schmoranzer as well as the referees for giving me helpful feedback. Also, I thank my fellow students and colleagues among the Effective Altruism movement for giving me input and motivation to engage with this topic.

REFERENCES

[1] N. Bostrom, A. Dafoe, C. Flynn. *Policy Desiderata in the Development of Machine Superintelligence*. (2016).
[2] P. Singer. *Animal Liberation.* HarperCollinsPublishers (2002).
[3] J. Bentham. *Introduction to the Principles of Morals and Legislation.* (1789).
[4] D. Chalmers. Mind Uploading: A Philosophical Analysis. In: *Intelligence Unbound: The Future of Uploaded and Machine Minds*. Wiley-Blackwell (2014).
[5] N. Bostrom, E. Yudowsky. The Ethics of Artificial Intelligence. In: *The Handbook of Artificial Intelligence.* Cambridge University Press (2014).
[6] J. Basl. Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines. In: *The Machine Question: AI, Ethics and Moral Responsibility.* The Society for the Study of Artificial Intelligence and Simulation of Behaviour (2012).
[7] D. Chalmers. Facing Up to the Problem of Consciousness. In: *The Character of Consciousness.* Oxford University Press (2010).
[8] N. Bostrom. Ethical Issues in Advanced Artificial Intelligence. In: *Science Fiction and Philosophy: From Time Travel to Superintelligence.* Wiley-Blackwell (2003).

[A] A. Chella, R. Manzotti. Artificial Consciousness. In: P*erception-Action Cycle: Models, Architectures, and Hardware.* Springer (2011).
[B] A. Mannino, D. Althaus, J. Erhardt, L. Gloor, A. Hutter, T. Metzinger,. Artificial Intelligence: Opportunities and Risks. Policy paper by the Effective Altruism Foundation. (2015).
[C] A. Reggia. The rise of machine consciousness: Studying consciousness with computational models. In: *Neural Networks*. Elsevier (2013).
[D] R. Manzotti. Machine Consciousness: A Modern Approach. In: *Natural Intelligence.* The INNS Magazine (2013).
[E] R. Shafer-Landau.: *Ethical Theory. An Anthology*. Wiley-Blackwell (2013).
[F] T. Metzinger. *Philosophie des Geistes - Band 1: Phänomenales Bewusstsein.* mentis (2006).